2. TEORI PENUNJANG

2.1. Tinjauan Pustaka

Bab ini menjelaskan mengenai teori-teori yang akan digunakan dalma penulisan skripsi dan pembuatan aplikasi.

2.1.1. Bahasa Bali

Bahasa Bali merupakan salah satu bahasa daerah Indonesia yang berasal dari pulau Bali. Bahasa Bali juga banyak digunakan oleh masyarakat di daerah Nusa Tenggara Barat, Jawa Timur, Lampung, dan Sulawesi Tenggara(*Bali - Peta Bahasa*, n.d.). Penggunaaan bahasa Bali dikalangan generasi Z masih berada di bawah Jawa Timur yang berada di angka 88,29% dan Nusa Tenggara Barat yang berada di angka 86,75%. Badan Pusat Statistik Provinsi Bali merilis persentase penggunaan bahasa Bali berdasarkan generasi. Generasi Post gen Z atau anak-anak berusia 9 tahun yang menggunakan bahasa bali sebanyak 78,82% dengan keluarga dan 75,89% dengan orang-orang di lingkungan sekitar. Ada juga generasi Z yaitu yang saat ini berusia 25-10 tahun sebanyak 88,07% dengan keluarga dan 84,20% dengan lingkungan sekitar. Penurunan persentase ini terjadi dari generasi *Pre-Boomer* ke generasi *Post Gen Z* yang mana merupakan akibat dari terjadinya pernikahan campuran antara suku Bali dengan suku lainnya. Juga pergaulan dengan tamu asing dan munculnya sekolah bertaraf internasional yang tidak memiliki mata pelajaran bahasa Bali menyebabkan rendahnya penggunaan bahasa Bali pada *Post Gen Z(Candrawati, 2023)*. Dalam penggunaan kehidupan sehari-hari, Bahasa bali secara tidak sadar sering sekali digunakan. Salah satunya pada daerah perkotaan, dimana dengan adanya proses modernisasi memperlihatkan hubungan yang sangat menonjol antara bahasa Bali dan etnisnya (Parwati & Sudiartha, 2021).

2.1.2. Transformer BERT

Bidirectional Encoder Representations atau yang sering dikenali sebagai *BERT* merupakan *pre-trained* dari *Bidirectional transformer* dari teks yang tidak memiliki label yang melihat kondisi dari semua layer yang berada dalam 1 corpus yang besar. Model ini telah dikembangkan oleh Google sejak tahun 2018. Fungsi dari metode *BERT* ini adalah untuk membantu model memahami bahasa yang ambigu. Dan keunggulan yang dimiliki oleh *BERT* dibandingkan dengan *language model* yang lain adalah *BERT* ini dapat melakukan pembacaan pada *Input text* dari dua arah yaitu

kiri kenan maupun kanan ke kiri. Transformer pada model *BERT* berfungsi untuk meningkatkan kapasitas pada model untuk memahami konteks dari suatu kata sehingga dapat membantu model memahami arti dari seluruh kalimat. *BERT* menyediakan beberapa *pre-trained* model yang telah dilatih dengan menggunakan data dari *Wikipedia* salah satunya adalah *BERT Multilingual.* Pada model ini memiliki kekurangan yaitu model ini tidak dapat mendeteksi satu bahasa atau pemilihan bahasa sehingga proses tokenisasi akan mecampur kata tersebut dari berbagai bahasa (Husin, 2023).

2.1.2.1. IndoBERT

IndoBERT adalah singkatan dari Indonesia Bidirectional Encoder Representations from Transformers. IndoBERT merupakan model bahasa alami terbaru yang dikembangkan oleh komunitas AI Indonesia yang memanfaatkan NLP dalam pengerjaannya. IndoBERT dalam penggunaan untuk menerjemahkan bahasa Indonesia ke bahasa lain memiliki akurasi yang tinggi (Andika, 2023). IndoBert merupakan versi Indonesia dari Transformer BERT. IndoBERT telah di *train* menggunakan lebih dari 220 Juta kata yang diambil dari 3 *main source* yang berbeda. 3 *main source* tersebut adalah Wikipedia Indonesia, artikel berita seperti Kompas, Tempo, dan Liputan6. IndoBERT digunakan untuk penggunaan IndoLEM, IndoLEM adalah *benchmark* Indonesia yang terdiri dari 7 tugas yang berbeda untuk bahasa Indonesia, *spanning morpho-syntax, semantics,* dan *discourse*(Koto et al., 2022).

2.1.3. Bi-LSTM

BI-LSTM merupakan LSTM yang dapat mengatasi masalah yang dimiliki oleh LSTM dimana Bi-LSTM dapat memproses kata dalam 2 arah yaitu *input forward* dan *input backward*(Alghifari et al., 2022). Pada BI-LSTM *training* dilakukan sebanyak 2 kali, dengan begitu BI-LSTM akan memberikan jaringan tambahan dan hasil yang didapatkan menjadi lebih lengkap.

Gambar 2.1. Cara kerja Bi-LSTM

Seperti contoh gambar diatas, terkadang kata yang akan kita gunakan tidak hanya ditujukan oleh kata sebelumnya, tetapi bisa juga kata setelahnya. Contohnya seperti gambar diatas, kata “Teddy” tidak berarti akan memunculkan kata “*bears*” atau “*Roosevelt*” tetapi bisa saja memiliki arti yang berbeda tergantung sebagaimana penggunaan kata itu diinginkan(Aggarwal, 2021).

2.2. Tinjauan Studi

2.2.1. Penerapan Convolutional Neural Networks untuk Mesin Penerjemah Bahasa Daerah Minangkabau Berbasis Gambar (Santoni et al., 2021)

Pada penelitian ini dilakukan penerjemahan terhadap gambar tulisan bahasa daerah Minangkabau dengan metode CNN. penelitian terkait mesin penerjemah khususnya bahasa Indonesia ke bahasa Daerah sudah banyak dilakukan. Penelitian yang dilakukan sebelumnya menggunakan dataset berupa teks sehingga, peneliti ingin merasa mesin penerjemah belum bekerja secara optimal. Disisi lain, banyaknya teks juga menjadi masalah karena akan menyulitkan pengguna. Untuk itu peneliti memilih untuk menggunakan gambar sebagai dataset dari penelitian karena kurangnya penerjemahan berbasis gambar/citra. Tujuan dari penelitian ini adalah, melestarikan bahasa daerah yang ada. Metode CNN dipilih karena CNN menggunakan *high level feature* yang mana gambar tidak perlu didefinisikan secara spesifik.

Pada tahap awal penelitian, peneliti melakukan studi literatur dengan mengumpulkan informasi, landasan teori, dan juga hasil dari penelitian sebelumnya. Tujuannya, peneliti ingin memperbanyak pengetahuan yang dimiliki agar dapat menemukan *state of the art* penelitian ini. Dari hasil penelitian sebelumnya didapatkan algoritma *Leveinstan Distance* merupakan algoritma yang memberikan akurasi yang paling baik dalam pemrosesan data teks.

Kemudian dilakukan pengumpulan data yang mana data dibagi menjadi 3 kelompok. Data pertama berisi karakter/alfabet/huruf kecil dan kapital dengan 5 font berbeda dan dari 5 *device* berbeda. Spesifikasi dari setiap *device* dapat dilihat dalam gambar berikut.

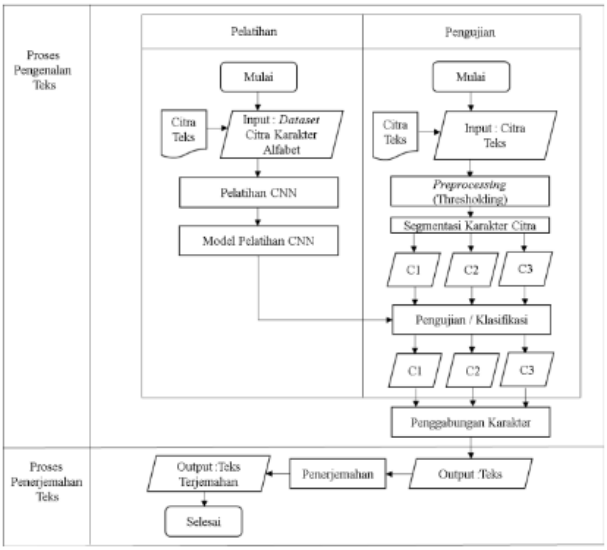
Tabel 1. Spesifikasi device/perangkat pengambilan citra

Jenis Perangkat (Kode)	Megapixel/ Scanner Type	Bukaan (f/)/ Sensor Type	Ukuran piksel / Optical Resolution
Scanner EPSON (D1)	600 x 1200 dpi	CIS	Flatbed Colour Image Scanner
Samsung Galaxy J5 (D2)	13 MP	f/1.9	1.15 µm
Iphone 4 (D3)	15 MP	f/2.8	1.5 µm
Iphone 5S (D4)	8 MP	f/2.2	1.5 µm
Iphone 7 (D5)	12 MP	f/1.8	1.22 µm

Gambar 2.2. Gambar Tabel spesifikasi pengambilan citra

Dataset kedua, terdapat juga data alfabet yang terdiri atas kata yang akan digunakan untuk menguji model CNN yang telah dihasilkan sebelumnya. Dataset ketiga merupakan kamus bahasa Minangkabau beserta artinya dalam bahasa Indonesia.

Setelah penelitian diperoleh, akan dilakukan perancangan model penelitian. Akan dilakukan penentuan terhadap constraint penelitian seperti, jumlah data, ukuran gambar, dan arsitektur penelitian yang akan digunakan. Kemudian dilakukan *training* untuk mendapatkan model CNN yang dapat mengenali seluruh alfabet. Selanjutnya dari data yang didapatkan akan digunakan untuk *training* dengan metode CNN, sehingga diperoleh *training model CNN*.



Gambar 2. Perancangan Eksperimen Penelitian

Gambar 2.3. Perancangan eksperimen Penelitian

Tahapan kedua adalah *testing* untuk menguji model CNN yang telah didapatkan dari tahapan sebelumnya. Pada tahap ini, *dataset* yang digunakan berupa gambar teks berisikan kata bahasa Indonesia. Hasil segmentasi berupa alfabet dari kata teks yang terdapat dalam setiap

gambar. Selanjutnya setiap karakter tersebut akan diujikan ke dalam model *training CNN*. hasil klasifikasi akan digabungkan kembali untuk membentuk kata yang nantinya akan diterjemahkan kedalam bahasa daerah. Kemudian dilakukan evaluasi terhadap model yang telah dibuat.

Hasil dari penelitian ini adalah akurasi dari model CNN untuk keseluruhan alfabet diperoleh sebanyak 98.97% dengan mayoritas alfabet dapat dikenali sebesar 100% namun tidak untuk huruf "f, i, t, v". Hasil yang kedua adalah akurasi klasifikasi teks dimana teks CNN akan digunakan untuk mengidentifikasi teks pada dataset teks yang mendapatkan akurasi sebesar 50.72%. Dan hasil ketiga adalah akurasi terjemahan, dimana teks bahasa Indonesia akan diterjemahkan ke bahasa Minangkabau dengan menggunakan algoritma *Leveinstan Distance*. Didapatkan tingkat akurasi sebesar 75.78% dimana algoritma *Leveinstan Distance* dapat memperbaiki kesalahan dalam proses identifikasi sebelumnya. Kekurangan dari penelitian ini adalah kurangnya tingkat akurasi sehingga beberapa kata masih sulit untuk dikenali.

2.2.2. Pendekatan Metode Transformers untuk Deteksi Bahasa Kasar dalam Komentar Berita Online Indonesia (Rendragraha, 2021)

Pada penelitian ini dilakukan deteksi terhadap komentar terhadap berita online di Indonesia yang berisi bahasa kasar dengan metode Transformer BERT. komentar kasar dapat memberikan pengaruh negatif bagi semua orang. Untuk itu diperlukan sistem yang dapat mendeteksi bahasa kasar dalam suatu kalimat. Yang mana salah satunya dilakukan dengan metode klasifikasi teks tujuan dari penelitian ini adalah mengimplementasikan dan mengevaluasi keakuratan dalam klasifikasi kalimat. Peneliti memilih menggunakan metode Transformer BERT karena metode ini mengatasi permasalahan terhadap klasifikasi teks dan mempengaruhi hasil penelitian. Metode yang digunakan dalam evaluasi adalah nilai Macro Average F1-Score dan nilai F1-score untuk masing-masing kelas label.

Penelitian ini menggunakan 2 dataset yaitu, dataset pertama digunakan untuk membuat *pretrained model* dan dataset kedua digunakan untuk melakukan *fine-tuning* untuk klasifikasi teks. Dataset untuk *pretrained model* akan berisi banyak kalimat tanpa adanya label yang didapat dengan melakukan metode *crawling* pada twitter dengan *keyword* kata kasar. Dataset kedua didapatkan dari portal-portal berita yang telah dibeli label. Komentar dipilih berdasarkan trend mulai dari bulan maret 2019 hingga september 2019.

Kemudian akan dilakukan pembangunan model untuk BERT dengan menggunakan *library* transformer. Kemudian dilakukan *preprocessing* dengan melakukan tokenizer. Kemudian *fine-tuning* dilakukan untuk klasifikasi teks dengan membangun *library ktrain.* Kemudian dilakukan

evaluasi untuk mengukur performa dari model yang telah dibangun. Hasil yang diperoleh adalah seperti tabel berikut.

Tabel 5. Hasil Pengujian

Model - Dataset	Macro Average-F1	F1 Non Offensive	F1 Normal	F1 Offensive
scratch - Langsung	49%	23%	93%	32%
scratch - R. Undersampling 50%	50%	15%	92%	43%
scratch - R. Undersampling 45%	48%	15%	91%	36%
scratch - Oversampling (nlpaug)	45%	7%	92%	35%
BertMulti - Langsung	54%	11%	95%	57%
BertMulti - R. Undersampling 50%	53%	11%	93%	55%
BertMulti - R. Undersampling 45%	51%	14%	93%	47%
BertMulti - Oversampling (nlpaug)	44%	17%	94%	20%
Random Forest - Langsung Random Forest	32%	0%	92%	5%
- R. Undersampling 50% Random Forest -	38%	15%	90%	8%
R. Undersampling 45% Random Forest -	31%	0%	91%	3%
Oversampling (nlpaug)	30%	3%	70%	15%

Gambar 2.4. Tabel Hasil Pengujian

Dapat dilihat dari tabel diatas, bahwa hasil dari *Random Forest* kurang baik dibandingkan dengan metode BERT. Label dengan performa paling tinggi didapatkan oleh label normal dengan hasil > 90%, diikuti oleh label *offensive*, dan yang paling rendah adalah label *offensive*. Secara keseluruhan, model BERT Multilanguage dengan nilai performansi Macro Average F1 untuk Scratch model adalah 50% pada dataset random undersampling 50% dan untuk BERT Multilanguage adalah 54% pada dataset langsung. Kekurangan dari penelitian ini adalah kurangnya dataset dan pemahaman yang dimiliki.

2.2.3. Pembuatan Aplikasi Mesin Penerjemahan Menggunakan Metode No Language Left Behind dari bahasa Indonesia ke bahasa Banjar (Pranata & Nurhidayat, 2023

Pada Penelitian ini dilakukan pembuatan mesin penerjemahan dari bahasa Indonesia ke bahasa Banjar dengan metode *No Language Left Behind*. Penelitian ini memiliki tujuan untuk dapat melestarikan bahasa banjar di Indonesia. Metode ini merupakan model kecerdasan buatan yang mampu membuat model penerjemah dengan sumber bahasa sebanyak 200 bahasa, keunggulan dari metode ini adalah metode ini dapat menerjemahkan banyak bahasa, termasuk bahasa-bahasa yang kurang umum, dan juga ini memiliki sifat *multilingual* yaitu dapat menangani beberapa bahasa secara bersamaan.

Dataset atau pengumpulan data yang digunakan untuk penelitian ini diambil melalui beberapa sumber yaitu data kamus dari internet, wawancara dengan narasumber orang dengan suku asli banjar, dan mengambil beberapa hasil terjemahan yang dihasilkan dari mesin penerjemah di internet. Implementasi aplikasi ini menggunakan bahasa pemrograman *python*

dan menggunakan *Fast API* sebagai *API* yang nantinya akan digunakan pada platform berbasis web dan *ReactJS* sebagai *framework front-end* untuk platform web.

Setelah mengumpulkan *dataset* akan dilakukan *preprocessing* terlebih dahulu, dari yang awalnya dalam bentuk *Microsoft excel* akan diubah menjadi *file* dengan format *csv.* Kemudian akan dilakukan penghapusan pada tanda baca yang salah, karena jika tidak dihapus maka akan muncul *error* karena tidak sesuai dengan format. Selanjutnya adalah melakukan pengujian dengan beberapa model penerjemah yaitu *Glosbe* dan *ChatGPT*, Pengujian model translasi dilakukan dalam beberapa tahap, termasuk membandingkan hasil terjemahan Bahasa Banjar ke Bahasa Inggris dan Indonesia serta uji kembali terjemahan tersebut ke dalam bahasa asalnya. Dilakukan juga pengujian melibatkan reference target untuk memperbaiki hasil terjemahan.

Hasil implementasi model pada API dan aplikasi web dievaluasi, menunjukkan waktu prediksi, hasil terjemahan, dan inferensi yang dihasilkan. Aplikasi website dikembangkan dengan fitur pilihan bahasa sumber dan target, menampilkan hasil terjemahan serta memberikan kemampuan untuk menukar target bahasa yang diterjemahkan. Dan waktu pengujian untuk satu kali penerjemahan adalah 9,40 detik untuk prediksi dan 9,53 detik untuk mengakses API.

2.2.4. Translation of the Lampung Language Text Dialect of Nyointo the Indonesian Language with DMT and SMT Approach (Abidin et al., 2021)

Pada penelitian ini melakukan penerjemahan dari teks bahasa Lampung dialek Nyo ke bahasa Indonesia yang dilakukan dengan 2 metode yaitu, *Direct machine translation* (DMT) dan *Statistical Machine Translation* (SMT). penelitian ini bertujuan untuk membantu siswa/i pendatang di Lampung dalam menerjemahkan bahasa Lampung dialek Nyo melalui purwarupa atau model yang dibangun. Untuk menerjemahkan teks bahasa Lampung dialek Nyo bisa menggunakan kamus, tetapi dengan menggunakan kamus akan mengharuskan orang-orang untuk membuka kamus secara berulang kali. Untuk itu penelitian ini didasarkan pada kamus berbahasa Lampung dialek Nyo. pendekatan untuk membuat mesin penerjemah ini adalah menggunakan DMT dengan kamus, atau pendekatan *rule-based* dengan menggunakan serangkaian peraturan dalam bahasa, juga menggunakan banyak *corpus.*

Subjek dalam penelitian ini adalah Kamus bahasa Lampung dialek Nyo dan *parallel corpus* bahasa Lampung dialek Nyo ke bahasa Indonesia. Kamus yang digunakan merupakan kamus buatan Herman, S.Pd.I. *Parallel corpus* akan dibuat secara manual dengan media *notepad*. Pada penelitian ini juga akan menggunakan algoritma DMT yang dibuat untuk menerjemahkan bahasa Lampung dialek Nyo ke bahasa Indonesia dengan berbagai macam bentuk teks, baik itu kata,

kalimat, maupun paragraf. SMT akan digunakan untuk *pre-processing* yang akan terdiri dari sentence alignment, tokenization, cleaning, dan lowercase filtering.

Akan dilakukan pengambilan data untuk mencoba penerjemahan. Data akan diambil secara acak diantara bahasa Lampung dialek Nyo yang mana telah diterjemahkan oleh narasumber. Nantinya hasil dari percobaan ini akan dibandingkan dengan data awal menggunakan *e Bilingual Evaluation Understudy* (BLEU). BLEU adalah algoritma yang ditujukan untuk melihat kualitas dari hasil penerjemahan yang telah dilakukan oleh mesin. Hasil yang diperoleh dapat dilihat dari gambar dibawah.

Table 1. BLEU RESULTS ACCURACY VALUE IN DMT AND SMT

Translation Results	BLEU Value (%)	
	DMT	*SMT*
Lampung language Nyo dialect to Indonesian	39.32	59.85

Gambar 2.5. Hasil dari Akurasi Terbaik DMT dan SMT

Gambar diatas menunjukkan SMT memberikan akurasi yang lebih tinggi dibandingkan dengan DMT. Hal ini dapat terjadi karena SMT bisa melakukan pendekatan terhadap *training* data yang ada melalui *parallel corpus* dan *mono corpus* yang digunakan.

Hasil dari penelitian ini adalah akurasi dari DMT bisa sampai dengan 39.32% dan SMT sebesar 59.85%. DMT berguna apabila menerjemahkan kata yang ada didalam *database* tetapi tidak bisa membaca makna dibalik kata tersebut. Sedangkan SMT bisa belajar melalui *training* data yang diberikan serta bisa mempelajari makna dibalik kata tersebut.